Detection of Cyber-Physical Faults and Intrusions from Physical Correlations

Andrey Y. Lokhov*†, Nathan Lemons†, Thomas C. McAndrew‡, Aric Hagberg† and Scott Backhaus§
*Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
†Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545
‡Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405
§Materials Physics and Applications Division, Los Alamos National Laboratory, Los Alamos, NM 87545

Abstract—**Cyber-physical systems are critical infrastructures that are crucial both to the reliable delivery of resources such as energy, and to the stable functioning of automatic and control architectures. These systems are composed of interdependent physical, control and communications networks described by disparate mathematical models creating scientific challenges that go well beyond the modeling and analysis of the individual networks. A key challenge in cyber-physical defense is a fast online detection and localization of faults and intrusions without prior knowledge of the failure type. We describe a set of techniques for the efficient identification of faults from correlations in physical signals, assuming only a minimal amount of available system information. The performance of our detection method is illustrated on data collected from a large building automation system.**

1. Introduction

Cyber-physical systems are physical networks, governed by the laws of physics, but regulated by a control system coupled to computer networks that transmit the information required to optimize and control the physical networks for reliability and efficiency [1], [2]. Examples include, but are not limited to, smart grids, gas pipelines, civil infrastructures, autonomous automotive systems, automatic pilot avionics and process control systems. The interdependence of the cyber and physical networks makes the combined system more vulnerable to attacks; manipulation of the computer control network can leverage cyber-physical capabilities to cause damage or significantly degrade the performance of the critical infrastructure [3], [4].

The ability to detect and localize failures or attacks represents an important step towards the design of resilient cyber-physical networks and strategies for implementation of certificates for proportional response. It is natural to expect that indications of intrusion or misbehavior in the cyber subsystem are present as anomalies in the physical network. This fact can be used for searching for outliers in the data streams collected by the sensors monitoring the state of the physical system – a well-studied problem in a wide range of application domains [5]. Although anomalous changes in individual signals can be an indication of a major failure or a crude attack, they do not capture more sophisticated scenarios of coordinated intrusions. Therefore, it is important to take into account information from the spatiotemporal correlations of anomalies of individual signals. Exploiting these correlations might enable probabilistic localization of the intruder or failure within the network, and hence serve as a basis for building a proper response.

We study the problem of detection and localization of disturbances based on the analysis of spatiotemporal correlations between physical data streams. Our goal is to develop efficient methods for the detection and localization of failures within the cyber-physical system without reference to a predefined attack vector. Failure events can be very diverse, while attacks become more and more creative and sophisticated, so the detection methodologies cannot be based on scripted scenarios. In addition, detection methodologies which do not exploit prior knowledge of the topology of the physical network will have a broader range of application. Therefore, we deliberately do not incorporate any specific aspects of the physical system architecture in the algorithm design. Further desired requirements for detection and localization algorithms include scalability (the number of signals and time measurements can potentially be very large), generality (the signals are heterogeneous and of diverse nature), robustness (the signals can be noisy and incomplete) and low computational complexity (to allow deployment of the algorithm in a fast online fashion).

Cyber-physical intrusion detection and response methodologies will improve at much faster rates when the development and refinement is closely coupled with real-world experimentation that validates strengths and reveals weaknesses. The simplicity and generality of the detection algorithms are very important since they will allow for deployment in different cyber-physical systems. In this paper, we test our techniques on specific real-world data from an automated HVAC system in a large office building at Los Alamos National Laboratory (LANL).

We present a general protocol for detection and localization of disturbance which meet most of the aforementioned requirements. First, we develop a simple procedure for constructing a special correlation matrix out of detrended heterogeneous signals, making some assumptions on the anomaly signature we would like to be able to capture. Then, we use the correlation matrix to solve three crucial tasks: i) detection of the anomaly using spectral methods;

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.111

ii) localization of a subset of anomalous nodes within the system using low-rank approximations and biclustering methods; iii) finally, identification of the functional role of the inferred anomaly based on the sensor labels. We validate our framework on experimental real-world data collected from a building automation system at LANL.

2. Time Series Analysis and Correlation Matrix Construction

We consider the problem involving data from N physical sensors indexed by V. For each sensor $i \in V$ we are given a time series $X_i(t)$ collected at times $t \in T$. The data $X_i(t)$ can be heterogeneous real or integer valued signals and provides a (partial) description of a system. We assume that the spatial and temporal relationships between the sensors are unknown, but that we do have access to sensor labels. We also assume that the fluctuations of each time series in the system around their mean behavior during normal operations are essentially independent; this assumption is correct if the sensors are weakly coupled.

Formally, we say that during normal operations the observations $X_i(t)$ can be modeled as

$$X_i(t) = Y_i(t) + N_i(t) + S_i(t), \qquad (1)$$

where the $N_i(t)$ represent the uncorrelated random noise, $S_i(t)$ is a potential signal of attack or failure (correlated between sensors) which is absent during normal operations, and $Y_i(t)$, which we call the trace, describes the idealized operation of the system without noise. When the system is attacked or experiences a fault the affected parts of the system $U \subset V$ are expected to move away from the trace, $S_i(t) \neq 0$ for $i \in U$. We are interested in those cases when the signal is nonzero for a significant subset of sensors. It may occur that the signal-to-noise ratio is sufficiently low so that for each individual sensor the failure signal is not directly observable, but that it can be detected and becomes statistically significant when the subset of affected sensors are taken into account collectively. In these cases, the differences between the trace and the corresponding observations will become related. In other words, since the $S_i(t)$ values corresponding to a particular disturbance event are likely to be correlated, we expect that the correlation relations will become apparent in the detrended signals $X_i(t) - Y_i(t)$ if the signal (e.g. attack or failure) occurs at $t = \tau$ and lasts for T time steps. Our goal is to construct a suitable correlation matrix out of these time series which will enable the detection and localization of the undesirable changes in system state.

2.1. Detrending the Signals

Unfortunately, the traces $Y_i(t)$ are *a priori* unknown. In some cases they can be learned from an ensemble of repeating operations under normal behavior, but here we assume that this data might be unavailable. Thus we approximate the traces with a running mean, $\bar{X}_i(t) := \frac{1}{\tau_{\text{av}}} \sum_{t'=t-\tau_{\text{av}}/2}^{t+\tau_{\text{av}}/2} X_i(t')$, centered at t. This is a reasonable assumption if the traces Y_i are fairly smooth; however, this will not be a good assumption if the system changes modes of operation or otherwise undergoes rapid changes within the interval $[t - \tau_{\text{av}}/2, t + \tau_{\text{av}}/2]$.

Note that although the use of the centered running mean requires the knowledge of the signal in the future, we found that it produces better results with respect to the approach where the trailing mean is employed. At the same time, an online detection algorithm based on the centered mean will have a time-lag of $\tau_{\text{av}}/2$. There is hence a trade off between the quality of approximation and the speed of detection.

It seems intuitive that the choice of smaller τ_{av} would introduce a smaller time-lag, and thus would lead to better results. On the other hand, τ_{av} should be large enough to average out the small fluctuations caused by the terms $N_i(t)$. A similar argument implies that τ_{av} should be chosen to be close in size to the expected duration of an attack or fault signal one would like to be able to detect: if τ_{av} is much larger than this scale, the signal will be likely to be averaged out. In practice, there is often a range of reasonable choices for the length τ_{av} of the sliding window; one should choose the one which satisfies the requirements on a desired maximum time-lag of detection.

2.2. Construction of the Correlation Matrix

We calculate correlation matrices from the residuals (an example is depicted in Figure 1) of the detrended data streams $R_i(t) := X_i(t) - \bar{X}_i(t)$. At this point, one more parameter, the time interval τ_{corr} over which correlations are calculated, must be chosen. Ideally, this time window should be at least as large as the duration of the event we would like to detect. This time length, in general, is application dependent; typically, we are interested in the time scales which are a low multiple of τ_{av}. Thus, we calculate the standard Pearson correlation coefficient $\xi_{ij}(t)$ for each pair of residuals $R_i(t)$, $R_j(t)$ over the correlation window $[t - \tau_{\text{corr}}, t]$ of length τ_{corr}. This gives us the desired correlation matrix $M_{ij}(t) = \xi_{ij}(t)$ at each time instance. We are not interested in detecting the self-correlations which are trivially equal to one, so we put by definition $\xi_{ii}(t) = 0$ $\forall i \in$With our setup under normal operations, when the data streams can be modeled as in Equation (1) with $S_i(t) = 0$, we expect the detrended data streams to be uncorrelated, $\forall i \neq j, \;\; \mathbb{E}[\xi_{ij}(t)] = 0$. However, during an attack or failure we expect there to be a set of sensors $U \subset V$ such that $S_i(t) \neq 0$ for $i \in U$, and hence

$$\forall r \neq s, \; r, s \in U, \;\; \mathbb{E}[\xi_{rs}(t)] = \sigma_{rs} > 0, \qquad (2)$$

since the non-zero signals $S_i(t), \; i \in U$ of the attack are supposed to have a similar behavior.

3. Detection and Localization of Anomalous Submatrix

In this section, we present a protocol for detecting and localizing a group of anomalously behaving devices

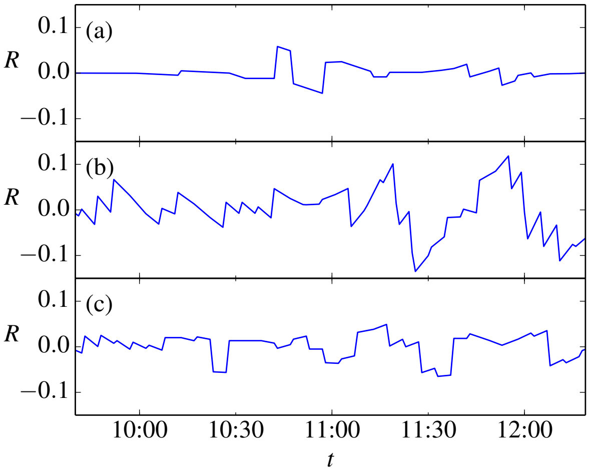

Figure 1. Three residuals R from a typical signal stream. Signal (a) has $S(t) = 0$ while signals (b) and (c) have correlated $S(t) \neq 0$ due to an attack or failure. The attack starts at approximatly 11:00 and some correlation can be observed between (b) and (c). The goal is to find and identify such correlated signals among the many recorded signals.

within the physical network. Formulating the problem in the framework of submatrix localization, the detection step is done by monitoring the spectral gap in the correlation matrix spectrum. This method is universal and does not require any prior assumptions on the form of the noise and on particular normalization of the correlation matrix. We explore three approaches to the localization of the anomalous nodes: sparse PCA based on a low-rank approximation, and two biclustering methods for finding a submatrix with an elevated mean value.

3.1. Detection of Anomalous Submatrix

Under normal conditions and low noise, the correlation matrix of the physical system might contain some structural information about the topology of the system. For instance, we can expect communities representing common functional roles or spatial locations of devices to have strong correlations. All other matrix elements should appear as noisy and uncorrelated values fluctuating around zero. When an anomaly occurs under the assumptions of Section 2 with a strong enough signal, one should witness the emergence of one single submatrix with a higher mean value. As in the problem of detecting a single community in a graph [6], the change in the correlation matrix induced by the anomalous signal should be also visible in the spectrum of the correlation matrix. In the ideal case, if the community is large enough, there is a spectral gap between the first and the second largest eigenvalues, and in addition, the principle eigenvector contains information about the location of the community. We use the idealized case to gain intuition about the behavior of the real world system.

This intuition for the correlation matrices constructed from the real signals comes from rigorous analysis for ideal noise, which also illustrates the concept of a "sufficiently strong signal" used above. As an example, consider a rank-1 matrix with eigenvalue θ, $P = \theta u u^T$, and suppose that we observe this matrix corrupted by a noise taking the form of a normalized $N \times N$ Gaussian Wigner matrix W, with zero-mean elements and variance of the off-diagonal elements equal to $1/N^2$. It is well known that the spectrum of W converges to the semi-circle law with support $[-2, 2]$. Let us denote the largest eigenvalue of the measurement matrix $P + W$ as λ_1, and the associated eigenvector as u_1. Depending on the "signal strength" θ, the values of λ_1 and u_1 undergo a phase transition [7]. If $\theta > 1$, then in the large N limit $\lambda_1 \to 1 + 1/\theta$ is clearly separated from the bulk, and $|\langle u, u_1 \rangle| \to 1 - 1/\theta^2$. In the opposite case $\theta \leq 1$, $\lambda_1 \to 2$ and the associated eigenvector does not carry any useful information, being completely degraded by the noise, with $|\langle u, u_1 \rangle| \to 0$. Similar results hold for the case of multiplicative noise.

The important question is how to decide whether the gap between the two largest eigenvalues $\Delta_1 = \lambda_1 - \lambda_2$ is statistically significant. The challenge here is that we do not assume any prior information on the statistics of the trace and on the noise distribution; this setting has not been well studied in the literature so far. To address this question, we suggest the following detection criterion. Let $\Delta_i = \lambda_i - \lambda_{i+1}$ be the collection of spacings between successive eigenvalues of the correlation matrix. Following the assumption that the nonzero values of all eigenvalues but the largest one are entirely due to a random noise, we can empirically estimate the corresponding characteristic noise scale as

$$\delta = \sqrt{\frac{1}{N-2} \sum_{1<i<N} \Delta_i^2}. \tag{3}$$

Now our proposed detection certificate is as follows: we consider that the first eigenvalue is statistically well separated if

$$\Delta_1 > \Delta_2 + \delta. \tag{4}$$

We count the opposite case as an absence of detection. The validity of this detection criterion will be checked in the Section 4 involving real data examples.

3.2. Localization Using the Low-rank Approximation

Once the detection certificate presented in Subsection 3.1 yields a positive result, the next step is to localize the anomalously correlated elements of the system. The K communities detection problem is often addressed using the low rank approximation [8]. In our case, a significant spectral gap Δ_1 indicates that the hidden matrix can be localized by looking at the best rank 1 approximation $\widehat{M}$ of the initial matrix M,

$$\widehat{M} = \arg\min_{\widehat{M}} \|M - \widehat{M}\|_F \quad \text{s.t. } \operatorname{rank}(\widehat{M}) = 1, \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm. The solution to this problem is well-known and is given by the singular value decomposition (SVD) of the matrix M, from which we retain only the leading singular value σ and the corresponding singular vector q [9]: $\widehat{M} = \sigma q q^T$. Unfortunately, in general the resulting vector q is not sparse, which does not allow us to identify the location of the anomalous nodes. Ideally, for detecting a group containing k anomalous nodes, we would like to obtain a vector with only k nonzero components, indicating their positions; this problem is often referred to as sparse PCA [10]. While under a general low-rank assumption this problem is NP-hard, for the special case of rank 1 it can be solved analytically simply by sorting the elements of q, and retaining only k largest elements [11], [12], resulting in a k-sparse vector that we denote as q_k. The constant in the expression for $\widehat{M}$ is then simply given by $\sigma_k = q_k^T M q_k$.

Another difficulty comes from the fact that *a priori* we do not know the size of the anomalous module. Sometimes, in order to find the optimal value of k, the so-called elbow method can be used [13]. The idea is fairly simple; find the minimal k such that the quality of approximation $\varepsilon_k \equiv \|M - \sigma_k q_k q_k^T\|_F$ is not increased "too much" when we make a step from k to $k+1$. More precisely, the optimal k is given by the minimal k such that $\varepsilon_k - \varepsilon_{k+1} < \epsilon$, where ϵ is some small constant, and the only parameter of the algorithm. The total complexity of the method is dominated by the complexity of the SVD-decomposition and is $O(N^3)$ in the most general case.

We expect the nonzero values of q_k for the optimal k to indicate the location of the nodes producing anomalous correlations. However, in the examples involving real data, the cusp on the elbow diagram might be not very pronounced in hard cases, therefore, in practice it can be unclear how to select an appropriate ϵ. At the same time it should be noted that at the end of the day we are not necessarily interested in inferring the whole set of anomalous nodes, but rather in understanding the cause of the anomaly. In this sense, one can choose to infer only a subset of anomalous sensors, but requiring a high level of confidence for this localization task; then the idea is to search for a subset of k^* strongly correlated nodes. However k^* can not be arbitrary small. Indeed, even in the idealized case there exist a practically achievable lower bound on the size of detectable community [14], [15] $k \gtrsim \sqrt{N}$. That is why the final suggested strategy consists in searching for a subset of most correlated sensors of size $k^* = \sqrt{N}$, and then in analyzing the corresponding group of devices using the tag data for determining the cause of the anomaly. This approach will be used in our experimental tests in Section 4, where an empirical evidence for the algorithmic failure in detection of communities of very small size will be presented.

3.3. Localization via Biclustering Methods

In this part we discuss two efficient algorithms for localization of the anomalous subgraph of the physical network, which do not explicitly use the rank 1 assumption, but instead attempt to find a $k \times k$ submatrix with an elevated mean. The first one, called Large Average Submatrix ($\mathcal{LAS}$), has been introduced in [16] and analyzed in Ref. [17], and consists in consecutive updates of k rows and k columns, starting from a random $k \times k$ submatrix and repeating the updates until a guaranteed convergence to a local maximum, meaning that the resulting submatrix can not be improved by changing only its column or row set. A recently introduced improved version of this algorithm, analysed in [18] and named Iterative Greedy Procedure ($\mathcal{IGP}$) follows a simple greedy scheme: starting by one randomly chosen row, we add the best columns and rows sequentially until a $k \times k$ submatrix is recovered. This algorithm outputs a provably better results, at least in the case of large Gaussian random matrices. In what follows, we test the performance of these algorithms on a real data set as a part of the localization procedure for finding the anomalously behaving group of nodes.

In order to get the best resulting submatrix, we use a multi-start procedure, initializing both algorithms L times for given k, and retain the most significant submatrix. As before, the size of the hidden subgraph k is unknown. In this case, again, we use $k^* = \sqrt{N}$ in order to find a smaller submatrix, representing the nodes which belong to the anomalous group of devices. The complexity of the overall algorithm based on this approach is dominated by the complexity of the localization step, and is equal to $O(N^3)$ for the low-rank algorithm, to $O(ILN \ln N)$ for $\mathcal{LAS}$ and to $O(2k^* LN \ln N)$ for $\mathcal{IGP}$, where I is the number of iterations needed for convergence of the $\mathcal{LAS}$ scheme ($I \lesssim 1000$ for practical cases described here), and $L \gtrsim 10^3$ is the number of warm starts that we use in biclustering algorithms to achieve a desired precision of the best local maximum.

3.4. Tests with synthetic data

We examined the detection procedure on artificially-generated signals consisting of a mixture of correlated and uncorrelated one-dimensional random walks. In this idealized situation, we generate $N = 900$ artificial signals as one-dimensional random walks starting from zero. We select $k_0 = 50$ of them to be correlated (repeating the step of the "master" random work with probability $\rho = 0.5$, and acting as independent otherwise) and to represent an anomalous subgroup we would like to detect and identify. First, we detrend the data and construct the correlation matrix M in the way described in Section 2; we choose $\tau_{\text{corr}} = 200$, and the running mean is taken over the window $\tau_{\text{av}} = 10$ time steps. The spectrum of M, shown in Figure 2, triggers a positive detection according to the criterion (4).

Next, we run the localization algorithms presented in Sections 3.2 and 3.3. We find that for $k^* = \sqrt{N} = 30$, all algorithms perfectly identify a subgroup of 30 correlated signals. If we choose to search the correlated group with the (unknown) ground truth size $k_0 = 50$, then the low-rank approximation approach misidentifies 5 signals, correctly counting the other 45 as correlated. Both biclustering methods make only one mistake in this case; however, it requires a rather large number of warm starts ($L \simeq 3 \cdot 10^4$) in order to converge to the best solution, which makes the

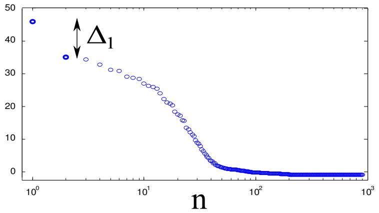

Figure 2. The spectrum (on a semi-log scale) of the correlation matrix M constructed from the total of $N = 900$ artificially-generated signals, including $k_0 = 50$ correlated walks. The correlated group of signals produces an identifiable gap Δ_1 in the eigenvalue spectrum.

algorithm slightly slower compared to the SVD-based one. As we will see in the next section, the speed of convergence is a very important property for online deployment of the algorithm.

4. Experiments with Real Data

4.1. System Description

Large commercial air conditioning (AC) systems represent an attractive cyber-physical test case for fault detection and localization algorithms because they contain relatively sophisticated physical, control and communications architectures, and the available tag data can serve as a ground truth for discovered groups and modules. We collected and analyzed the data streams from the AC system in a $30\,000\,\mathrm{m}^2$ office building, with about 900 sensors located in the conditioned spaces. These sensors record local temperature, airflow and valve opening positions. See Figure 3 (Left) for a schematic representation of the system used in this study, which shares a common structure with a large number of commercial AC systems. A more in-depth discussion of this AC layout is provided in [19]. Altogether this constitutes a system of approximately 1000 data heterogeneous data streams which are sampled once per minute. A network representation of a part of the physical system including conditioned spaces, fans and controllers is drawn in Figure 3 (Right); this data has been extracted from the tag data accompanying the recorded signals. This figure takes into account the spatial layout of conditioned rooms, and gives an idea of physical and communication links in the system.

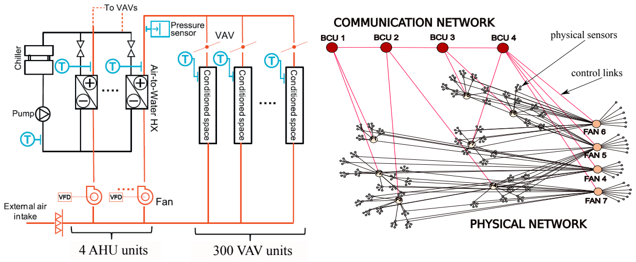

Figure 3. **Left:** A schematic representation of the air conditioning (AC) system. The recorded temperature, airflow and valve opening position signals from all the sensors and fans are used as input data streams. **Right:** Network representation of a part of the cyber-physical system which reflects the spatial organization of the conditioned spaces, and includes a part of both physical and control links.

A conflict of local control loops causes one fan (Fan 6 in Figure 3, Right) to behaving anomalously; at certain times of the day it produces mild uncontrolled oscillations. Although this action is not a result of an attack, it represents a perfect initial test for the protocol aiming at detection and localization of failures. We expect that these oscillations should leave a signature in the correlations of related physical signals even while the signal is too weak to be visible and identified as an outlier in individual recorded signals. This anomalous behavior is a proxy for attacks of the control architecture that can occur due to computer control network vulnerabilities. First we demonstrate the performance of our detection certificate using the described Fan 6 oscillations as an example failure event. Then we perform controlled experiments mimicking a simple intrusion on a smaller subset of devices in order to test the performance limits of the detection and localization algorithms as a function of the size of the anomalous set.

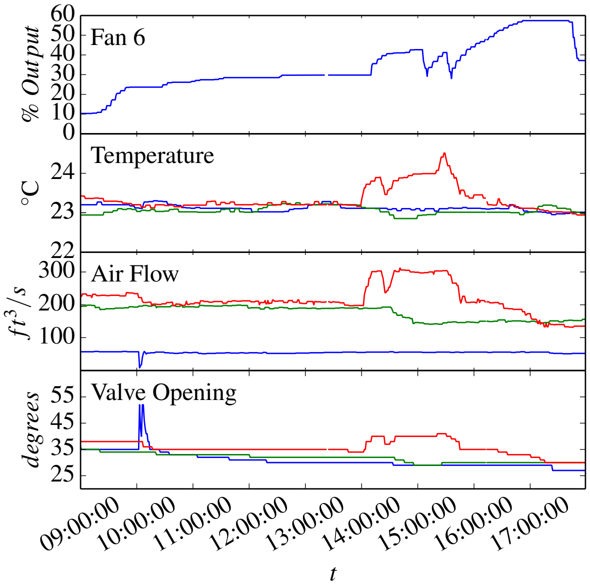

Figure 4. Fan 6 oscillations create anomalous data measurements in rooms that are serviced by that fan. Changes in output of Fan 6 (top plot) influences the temperature, air flow, and valve opening positions in Room 1 (red) and Room 2 (green) measurement data but not in Room 3 (blue) data, presented in the bottom plots: in this example, Rooms 1 and 2 are serviced by Fan 6 but Room 3 is not.

4.2. Detection Algorithm Performance

In Figure 4 we show examples of our data streams. The top plot of Figure 4 shows an anomalous behavior of Fan 6. The three bottom plots show examples of other signals of different types (temperature, airflow and valve positions) that we use for tests. The analysis of individual signals do not allow us to detect an anomalous behavior and to relate it to the malfunctioning Fan 6, and therefore we follow the procedure described in Section 2, constructing the correlation matrix and attempting to detect the anomaly from correlations of physical signals.

Let us first demonstrate the performance of the detection algorithm described in Section 3.1. In Figure 5, we show the spectra of the correlation matrices M in four different situations: i) Fan 6 oscillating, and all signals included; ii) Fan 6 oscillating, and signals serviced by Fan 6 removed from the data; iii) Fan 6 not oscillating, all signals included; iv) Fan 6 oscillating smoothly with a large period (on the order a half a day). It is clear that only case i) should trigger a positive detection outcome. Indeed, we notice that only the spectrum in this case satisfies the condition (4), while all other situations yield a negative detection result. The matrix M in each case has been constructed using the parameters $\tau_{\text{av}} = 30$ min and $\tau_{\text{corr}} = 200$ min.

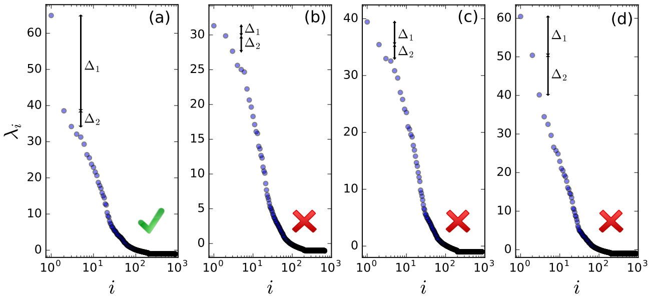

Figure 5. Spectra (in the semi-log scale) of the correlation matrix M for different scenarios. Oscillations of Fan 6 occur: (a) related signals included, (b) related signals excluded. All signals included when (c) Fan 6 does not oscillate and (d) Fan 6 oscillates, but with smoothly with a large period. Only the spectrum (a) satisfies the detection condition (4), as it should be.

4.3. Localization Algorithm Performance

Once the presence of anomaly is detected, we compare the performance of localization algorithms. Is it possible to correctly identify the group of nodes related to the anomalous fan, and hence to infer the reason of misbehavior? Tables 1 and 2 demonstrate localization results for two values of group sizes: the ground truth $k_0 = 209$ heterogeneous streams serviced by Fan 6 (out of $N = 974$ total signals), which is in general unknown, and for $k^* = 30$ strongest signals. We follow the strategy outlined in Sections 3.2 and 3.3 and use different combinations of the smoothing window time τ_{av} and the correlation time window τ_{corr}. As discussed in Section 2, little relevant information is captured with small τ_{av}, and indeed we find that $\tau_{\text{av}} = 10$ does not lead to a positive detection, see Table 1. The best results are obtained for larger values of τ_{av}, where more data is incorporated in the correlation matrix.

One of the major requirements for the algorithms is the ability to perform online detection and localization. New data points arrive every minute, so we would like the localization algorithms to converge in several seconds. The low-rank algorithm is very fast, and does not need any adjustments. As discussed in the previous section, in order to meet the computation complexity requirement for the biclustering algorithm we are forced to limit the number of warm starts to 1000 for the size $k_0 = 209$ and to 10000 for $k^* = 30$ since the convergence time of biclustering procedure grows with k. Another important property of the biclustering methods is that unlike in the low-rank approximation, the identities of the discovered columns do not always match the identity of the discovered rows; we use only one of the subsets to compute the number of mismatches.

TABLE 1. The number of mismatches (false detections) identified by the localization algorithms in the presence of Fan 6 activity for the searched groups of sizes k^* and k_0, with $k^* = 30$. For all cases, $\tau_{\text{corr}} = 120$ min is kept fixed.

τ_{av}	Detection	Number of false positives					
		Low-rank		$\mathcal{LAS}$		$\mathcal{IGP}$	
		k^*	k_0	k^*	k_0	k^*	k_0
10	✗	27	169	26	144	25	149
30	✓	0	123	0	112	0	115
50	✓	0	106	0	107	0	108

TABLE 2. Comparison of the localization algorithms under the same conditions as the ones described in Table 1, as a function of τ_{corr}. In this table, $\tau_{\text{av}} = 30$ min is kept fixed.

τ_{corr}	Detection	Number of false positives					
		Low-rank		$\mathcal{LAS}$		$\mathcal{IGP}$	
		k^*	k_0	k^*	k_0	k^*	k_0
90	✓	2	128	2	120	2	122
120	✓	0	123	0	112	0	115
160	✓	0	112	0	110	0	109
200	✓	0	106	0	103	0	104

With these restrictions, the three algorithms produce similar results with a comparable speed (under 3 seconds for low-rank algorithm and within $20 - 30$ seconds for biclustering in the present case). While only half of the true nodes are discovered when searching for all of the k_0 anomalous signals, very few false positives occur when only searching for the k^* strongest signals. The discovered k^* signals in almost all cases belong to a subgroup of a true group related to the anomalous fan. This value is sufficient to determine the common functional role of nodes inside this group, which corresponds to their relation to the anomalous Fan 6 in this case study. Therefore, all algorithms satisfy the requirements of performance, simplicity and scalability, which make them appropriate for deployment in real cyber-physical systems. In the next section, we discuss controlled experiments which would allow us to investigate the effect of the size of the anomalous community.

4.4. Identification limits from controlled experiments

Previously, we have tested the performance of the scheme on detecting the faulty behavior of Fan 6 already present in the system. In this section, we report results from controlled experiments on particular sensors of the office automation system. In their simplest form, these experiments consisted in a manipulation of temperature set points, mimicking localized intrusions of small amplitude. The trials were conducted on the controllers related to a small number of sensor units on Fan 5 (a non-oscillating fan, see Figure 3, Right), while all sensors related to the anomalous Fan 6 have been excluded to avoid an undesired interference.

The experiments that we report here took the following form: the temperature set points affecting 30 sensors related to Fan 5 and measuring temperature, airflow, and valve opening position were raised $0.5°F$ for 30 minutes and then lowered $1°F$ for the next 30 minutes. Among these potentially affected 30 data streams, only 16 showed a significant level of correlation. There are several reasons for this behavior, but the most important one consists in the observation that the airflow and valve opening positions have a much faster response to the set-point change compared to the temperature measurements which rise or fall on a much longer time scale. In the following we assume that these $k_0 = 16$ sensors constitute the ground truth for an anomalous group of nodes.

Using the collected data, we validate the choice of $k^* = \sqrt{N}$ put forward in Sections 3.2 and 3.3, and used throughout the study of the anomalous sensors related to the Fan 6. In particular, we verify that if the size of the group k_0 represents a sufficiently small fraction of the total number of signals, then it can not be correctly localized. In order to perform this study, we have considered 1000 selections of N randomly chosen signals but always containing the $k_0 = 16$ anomalous nodes. We applied our detection and localization protocol in each case for a range of N. The low-rank algorithm was used for localization as we have seen that at these scales it gives the same results with the fastest computation time; other localization methods show equivalent results. Note that the localization procedure was triggered only when the detection condition (4) was satisfied.

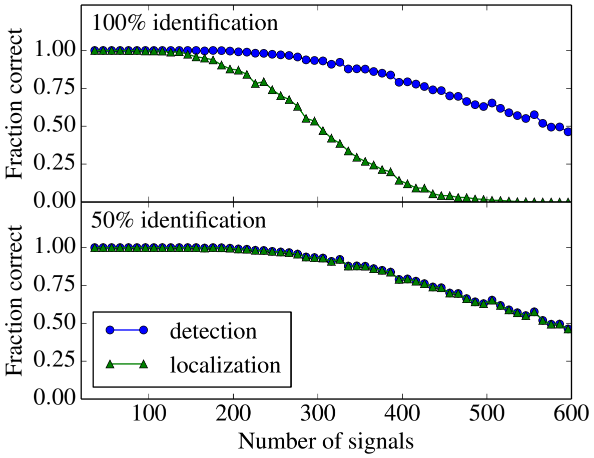

Figure 6. Empirical probability of successful detection and localization of a group of $k_0 = 16$ anomalous devices as a function of the total number of signals N. Localization is considered as succesfull if all k_0 nodes are correctly identified (top) and if at least 50% of nodes are recovered (bottom). Each point is averaged over 1000 random selections of N signals.

The results are presented in the Figure 6 with the empirical probability of successful detection and localization shown as a function of the total number of signals N. Two definitions of success are examined; a full and correct 100% identification of the ground truth, and a successful localization of at least 50% of the k_0 nodes, i.e. correctly identifying at least 8 devices out of 16. For the 100% identificaiton case we find a phase transition-type behavior as a function of N. The localization algorithm starts to fail at some point near $N = k_0^2$. This behavior is very close to the theoretical bounds derived in the idealized situations of Gaussian and Bernoulli distributions; in particular, it justifies our choice for k^* in the case where the optimal community size is unknown. The second case of 50% identificaiton illustrates that if we allow for some mistakes in the identification of anomalous sensors, then a successful localization occurs every time the detection procedure yields a positive result. This procedure might be appropriate if the labeled network is sufficiently sparse and the common cause of the anomaly can be easily identified using the sensor labels even in the case where not all the nodes are correctly localized.

5. Related Work

Methods for detecting and localizing cyber-physical failures and attacks have attracted significant attention [1], [2], [20], [21]. Major hurdles stem from a high degree of influence of sensor data from seasonal changes, proximity correlations and operational switches, and from the fact that infrastructure operators do not always have an accurate model of the physical network, or the existing models are not integrated into unified cyber-physical system model [20]. Another important factor is an increasing size and complexity of the systems under considerations. Some previous works develop detection techniques based on an accurate system modeling and on accounting for different attack scenarios [21], which represents an opposite approach to the problem compared to the present study which is agnostic to the specific aspects of the system architecture.

Aiming at general applications, we have used a simple running-mean signal detrending procedure in Section 2. Depending on a particular application, a wide array of other detrending methods [22], [23] can be used, each of them having associated strengths and weaknesses. The considered problem can be regarded as the detection of outliers which is an important field with application to a wide number of domains (see [24] for a survey). A large number of methods have been suggested, including network [25] and time series [5] specific techniques. A general formulation of the anomaly detection problem often takes form of hypothesis testing by considering H_0 (absence of anomaly) versus H_1 (presence of anomaly). In the present work, H_1 has been formulated as the existence of a submatrix with deviating elements in a properly normalized correlation matrix. This task is directly related to the problem of finding hidden cliques and community detection in graphs [6].

The detection of the anomalous submatrix is an instance of a problem known as optimal denoising which appears in many machine learning [26] and signal processing [27] applications. In real-world situations the signal matrix might have no special structure, while the form of the noise term is in general unknown. Several studies have explored the problem of the effective rank estimation of the signal matrix by optimal thresholding of singular values [28], [29]. In this work, we encountered a different problem of estimating the size of the anomalous submatrix under the fixed low-rank assumption.

6. Conclusions

We explored a set of methods for detection and localization of failures in cyber-physical systems which are based on the analysis of correlations between physical time series. The established protocol enables the identification of a group of anomalous sensors and provides insight for the localization of the failure source. The detection procedure achieves a number of important requirements, including low computational complexity and simplicity of implementation. Our capability to access the cyber-physical demonstration system to collect and analyze data from this system, and to deploy the detection algorithm opens a path forward for future work. We plan to continue real-world experiments which will consist of manipulating the building control system in a known manner using diverse attack strategies; this will allow us to further validate the presented methods. Another direction that we intend to explore consists of combining more control communication network data in order to minimize the possibility of false detections and to enhance the quality of failure source localization. These developments are essential for conception of algorithms for proportional response and for designing resilient cyber-physical networks.

Acknowledgments

The authors acknowledge Arthur Barnes, Gary Goddard and Hari Khalsa for their help with data collection, and Charles Bordenave, Michael Chertkov, David Gamarnik, Earl Lawrence, Sidhant Misra and N. Raj Rao for fruitful discussions. This work was funded by the Department of Energy at Los Alamos National Laboratory under contract DE-AC52-06NA25396 through the Laboratory-Directed Research and Development Program.

References

[1] L. Sha, S. Gopalakrishnan, X. Liu, and Q. Wang, "Cyber-physical systems: A new frontier," in *Machine Learning in Cyber Trust*. Springer, 2009, pp. 3–13.

[2] J. Shi, J. Wan, H. Yan, and H. Suo, "A survey of cyber-physical systems," in *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*. IEEE, 2011, pp. 1–6.

[3] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

[4] Y.-L. Huang, A. A. Cárdenas, S. Amin, Z.-S. Lin, H.-Y. Tsai, and S. Sastry, "Understanding the physical and economic consequences of attacks on control systems," *International Journal of Critical Infrastructure Protection*, vol. 2, no. 3, pp. 73–83, 2009.

[5] M. Gupta, J. Gao, C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 9, pp. 2250–2267, Sept 2014.

[6] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[7] F. Benaych-Georges and R. R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, vol. 227, no. 1, pp. 494–521, 2011.

[8] A. Coja-Oghlan, "Graph partitioning via adaptive spectral techniques," *Comb. Probab. Comp.*, vol. 19, no. 02, pp. 227–284, 2010.

[9] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[10] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.

[11] D. S. Papailiopoulos, A. G. Dimakis, and S. Korokythakis, "Sparse PCA through low-rank approximations," *JMLR: Workshop and Conference Proceedings*, vol. 28, no. 3, pp. 747–755, 2013.

[12] Z. Zhang, H. Zha, and H. Simon, "Low-rank approximations with sparse factors I: Basic algorithms and error analysis," *SIAM J. Matrix Anal. Appl.*, vol. 23, no. 3, pp. 706–727, 2002.

[13] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[14] B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," *arXiv preprint arXiv:1509.07859*, 2015.

[15] Y. Deshpande and A. Montanari, "Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time," *Foundations of Computational Mathematics*, vol. 15, no. 4, pp. 1069–1128, 2015.

[16] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, pp. 985–1012, 2009.

[17] S. Bhamidi, P. S. Dey, and A. B. Nobel, "Energy landscape for large average submatrix detection problems in Gaussian random matrices," *arXiv preprint arXiv:1211.2284*, 2012.

[18] D. Gamarnik and Q. Li, "Finding a large submatrix of a Gaussian random matrix," *arXiv preprint arXiv:1602.08529*, 2016.

[19] G. Goddard, J. Klose, and S. Backhaus, "Model development and identification for fast demand response in commercial HVAC systems," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 2084–2092, 2014.

[20] A. B. Sharma, F. Ivančić, A. Niculescu-Mizil, H. Chen, and G. Jiang, "Modeling and analytics for cyber-physical systems in the age of big data," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 74–77, 2014.

[21] R. Mitchell and I. R. Chen, "Modeling and analysis of attacks and counter defense mechanisms for cyber physical systems," *IEEE Transactions on Reliability*, vol. 65, no. 1, pp. 350–358, March 2016.

[22] D. R. Brillinger, *Time series: data analysis and theory*. SIAM, 2001, vol. 36.

[23] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[24] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[25] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 12, no. 2, pp. 159–170, 2010.

[26] R. Kannan and S. Vempala, *Spectral Algorithms*. Norwell, MA, USA: Now Publishers Inc., 2009.

[27] L. L. Scharf, "The svd and reduced rank signal processing," *Signal processing*, vol. 25, no. 2, pp. 113–133, 1991.

[28] R. R. Nadakuditi, "Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 3002–3018, 2014.

[29] S. Chatterjee *et al.*, "Matrix estimation by universal singular value thresholding," *Ann. Stat.*, vol. 43, no. 1, pp. 177–214, 2015.